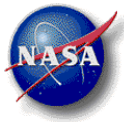
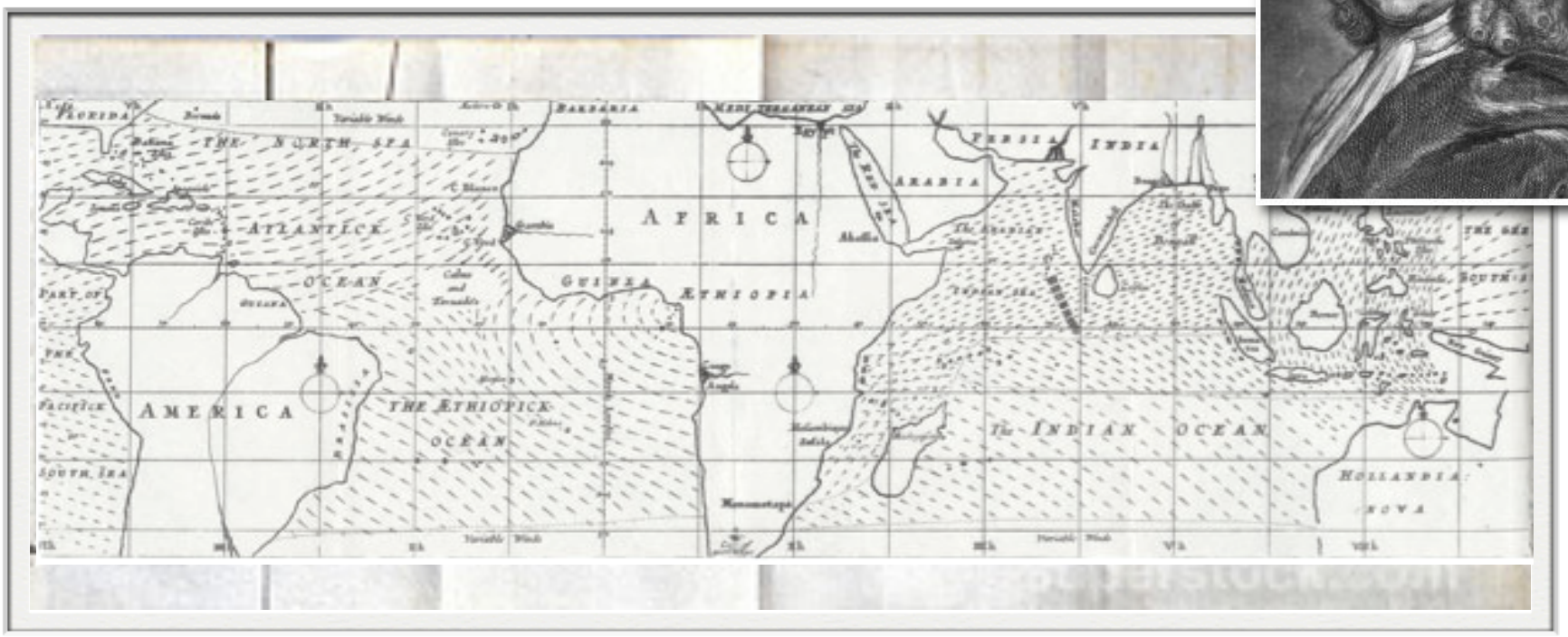


iRODS-Based Climate Data Services

John L. Schnase
Office of Computational and Information Science and Technology (Code 606)

NASA Goddard Space Flight Center
September 12, 2012

Climate Data and Climate Data Services



Edmund Halley's Wind Map (1686)
First global data image - a weather map, showing prevailing winds on a geographical map of the Earth.

Climate Data and Climate Data Services

The story of climate data

•It's old, global in scale, and all of it is relevant.

•It's one of the fastest growing classes of scientific data.

•Climate research is becoming increasingly data centric.

•Data publication and data sharing is becoming a big deal.

•The use of climate data in other domains is expanding.

Data services as a core mission

•Climate data services represent an effort to respond to the story of climate data.

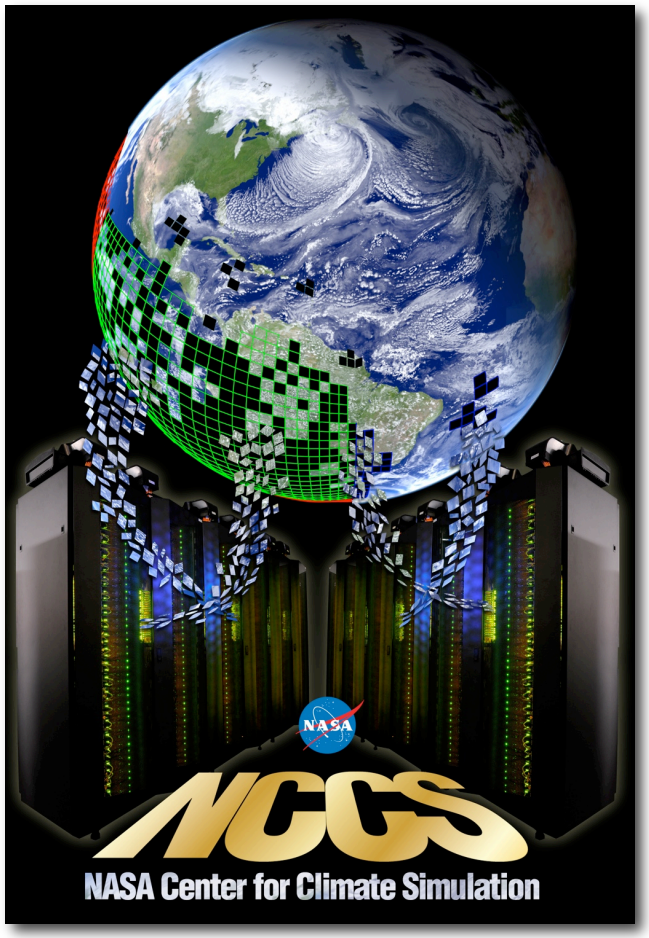
•The concept is being defined.

•At the very least, we know that it will require:

- a new view of data stewardship,

- new organizational processes, and

- new data technologies.



Climate Data and Climate Data Services

The story of climate data

•It's old, global in scale, and all of it is relevant.

•It's one of the fastest growing classes of scientific data.

•Climate research is becoming increasingly data centric.

•Data publication and data sharing is becoming a big deal.

•The use of climate data in other domains is expanding.

Data services as a core mission

•Climate data services represent an effort to respond to the story of climate data.

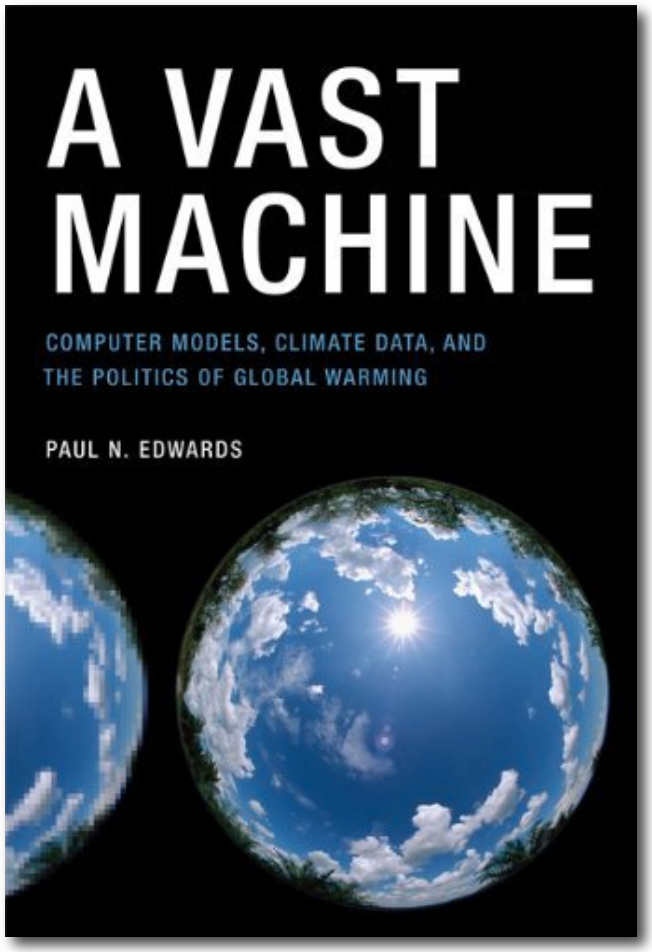
•The concept is being defined.

•At the very least, we know that it will require:

- a new view of data stewardship,

- new organizational processes, and

- new data technologies.



Canonical Scenario

Scenario

A customer approaches the NCCS with a new dataset they want us to manage ...

Q. What technology is needed to quickly meet that customer's requirement under the follow constraints:

- The solution should be: simple, fast, and cheap;
- provide core capabilities to get started, but extendable to accommodate future needs;
- be flexible, with the ability to use, optimize, and change deployment configurations in response to resource availability; and
- allow the new dataset to be integrated into an existing data collection?

We're looking at solutions that combine data grids and cloud computing ...

Definitions

Customer – an individual scientist, a lab, project, or mission.

Dataset – may be products generated by a climate model, may be observational data, reanalysis data, or specialized products.

Manage – may refer to short-term file storage, long-term archival preservation; data may be used online by a person or application.

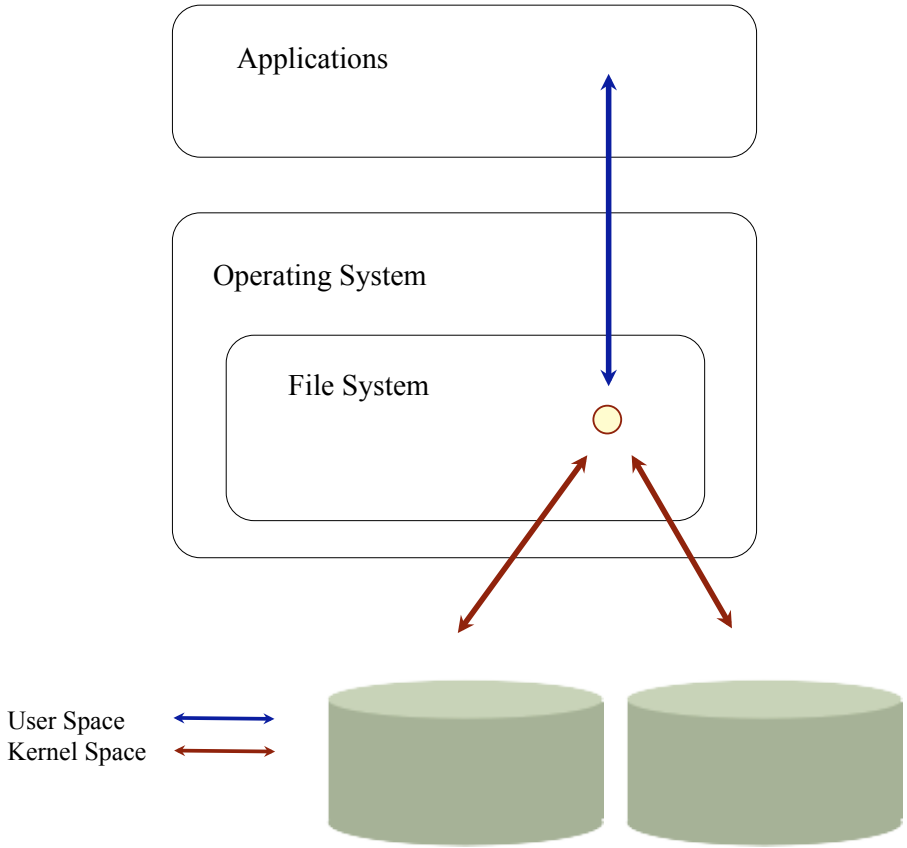
Examples

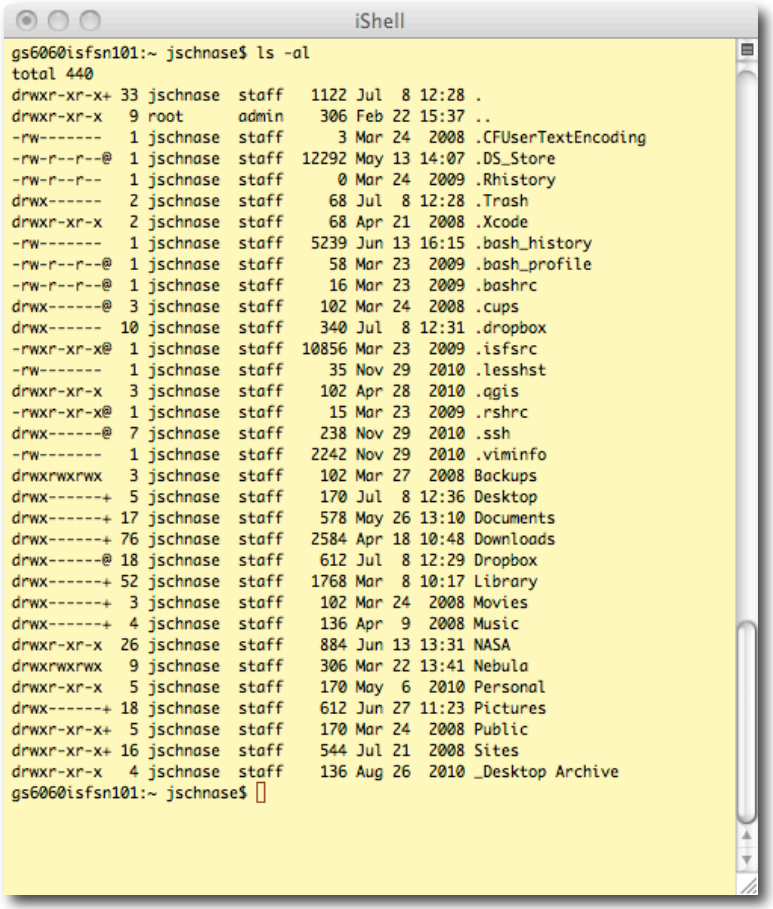
- IPCC AR5 data for ESGF.
- MODIS Atmospheres data for CMIP.
- AgMIP, SMOS, SMAP, TRMM, CERES, ...

Data Grid Software

POSIX Filesystem

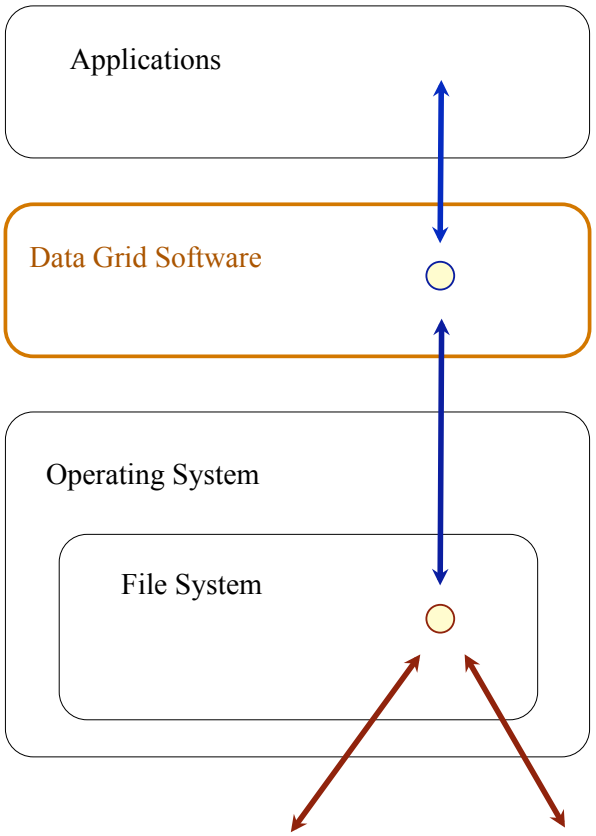
A set of standard interfaces to the Unix operating system. Traditional filesystems are typically “owned” by the system ...

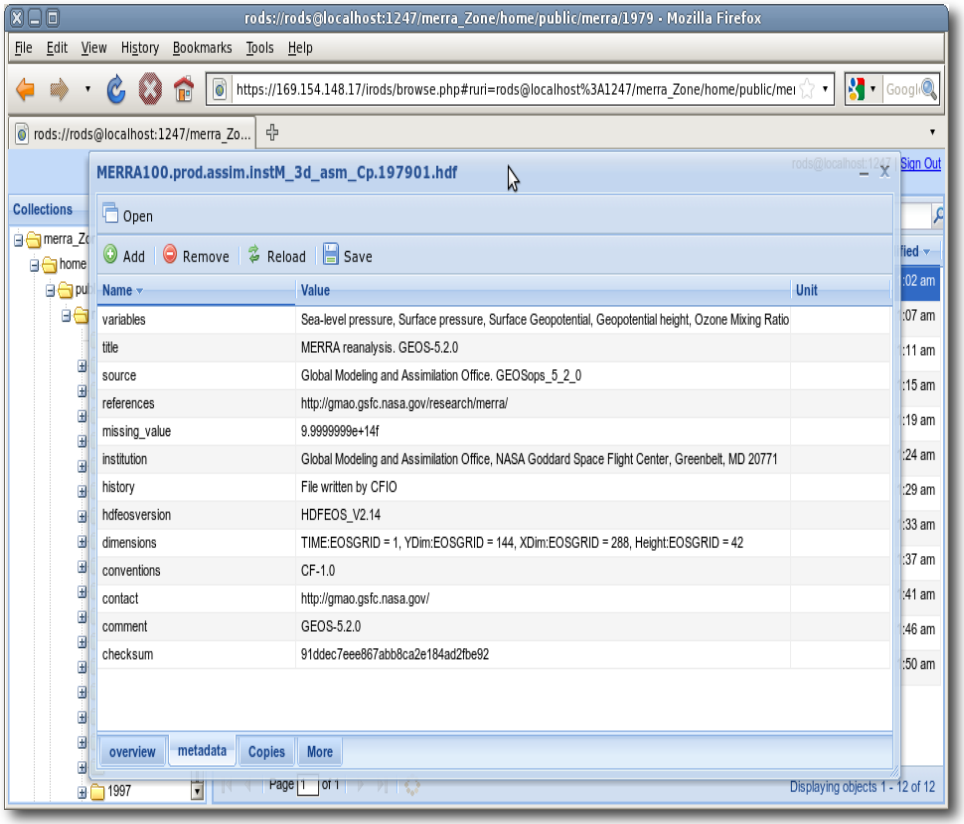




Standard POSIX filesystem ops and metadata over a traditional hierarchical filesystem ...

Data Grid Software

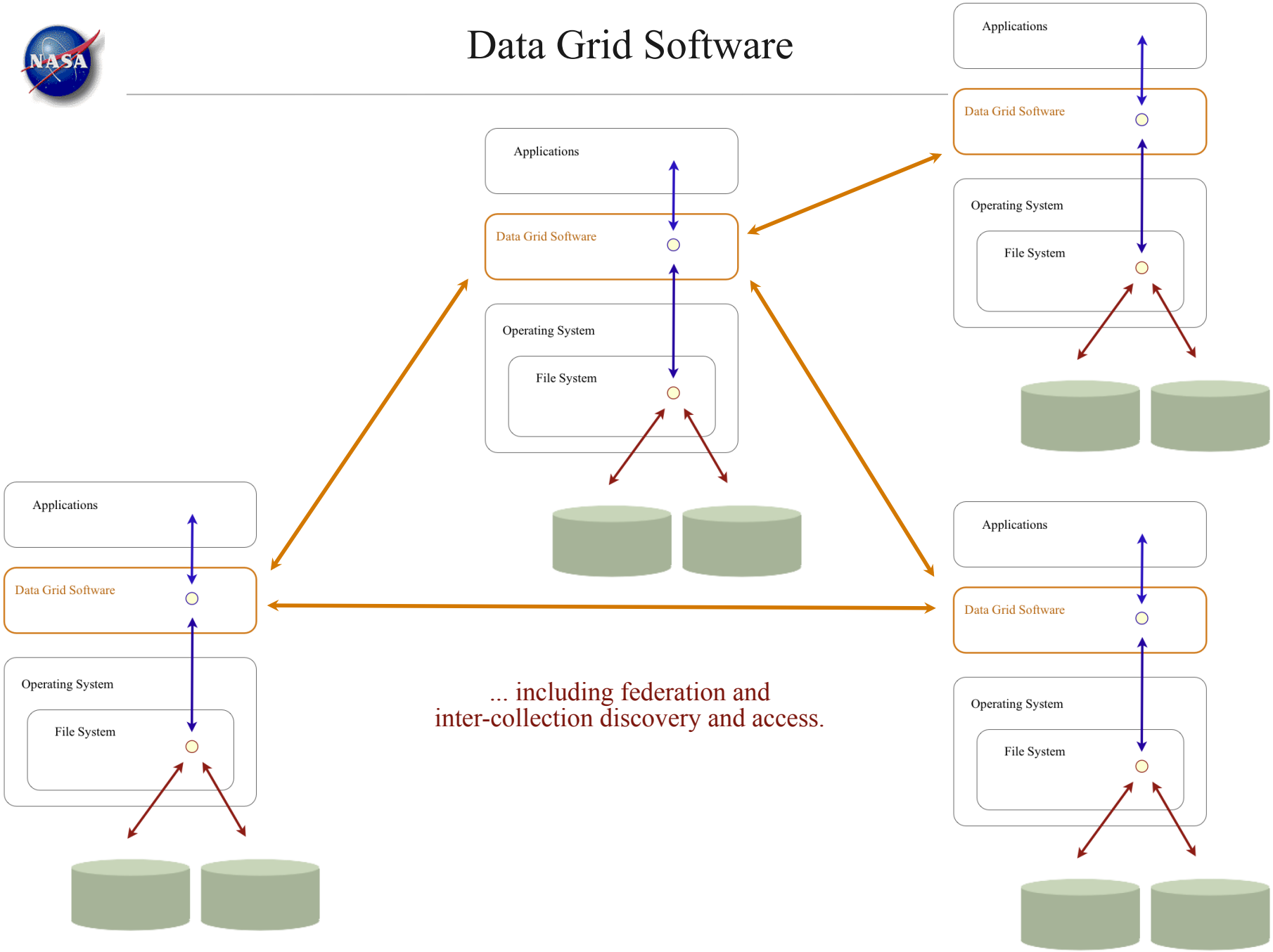




User Space
Kernel Space

Data grid "middleware" runs as an application in user space and provides a richer set of metadata descriptors and extended capabilities ...

Data Grid Software



... including federation and
inter-collection discovery and access.

iRODS: integrated Rule-Oriented Data System

Background

•Open source data grid software system.

•Developed by the Data Intensive Cyber Environments (DICE) group, University of North Carolina.

•Historic roots in data grids, digital libraries, persistent archives, and real-time data systems R&D, and SRB.

Features

•Targets large repositories, large data objects, digital preservation, and integrated complex processing.

•Supports server-side workflows implemented by chaining execution rules together based on data policies.

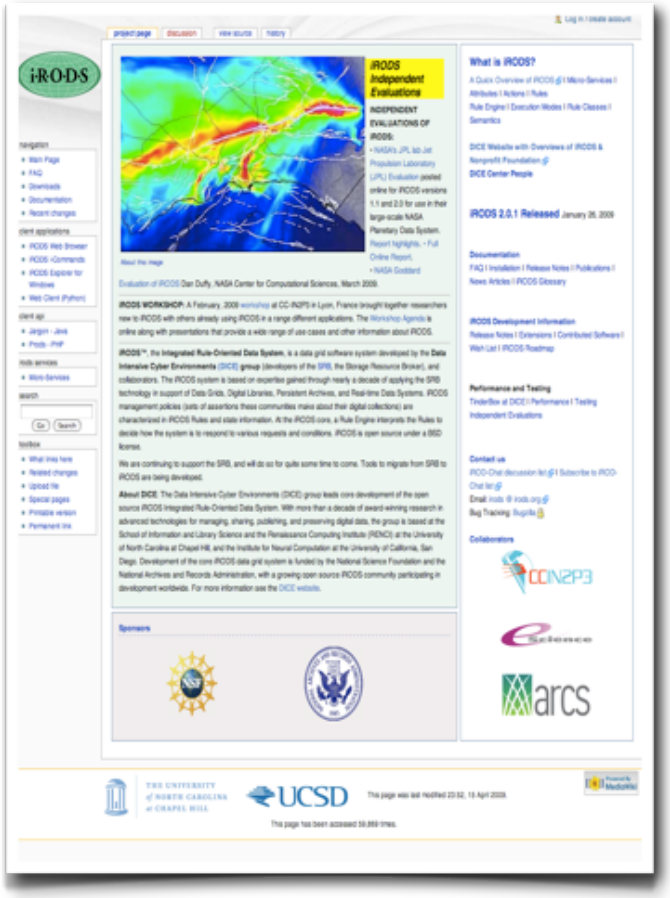
•Enables scalability and extensibility.

Major Concepts

•Policies => iRODS rules.

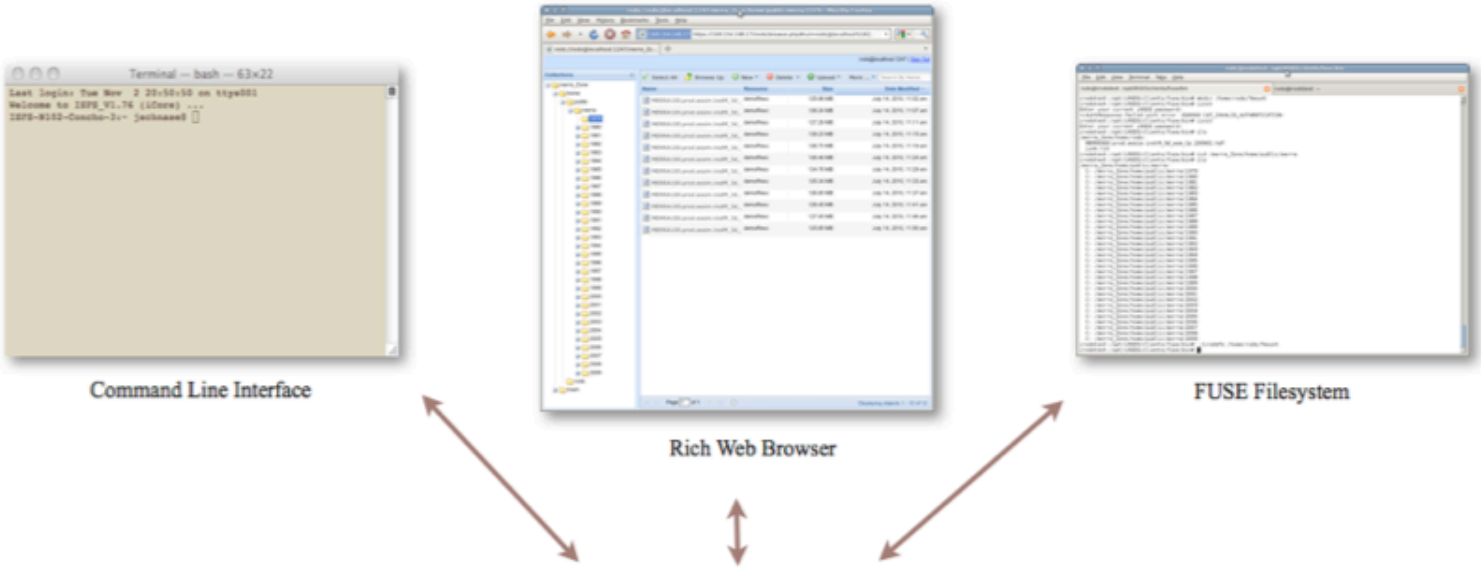
•Mechanisms => iRODS microservices.

With iRODS metadata providing the information necessary to perform these mappings ....



www.irods.org

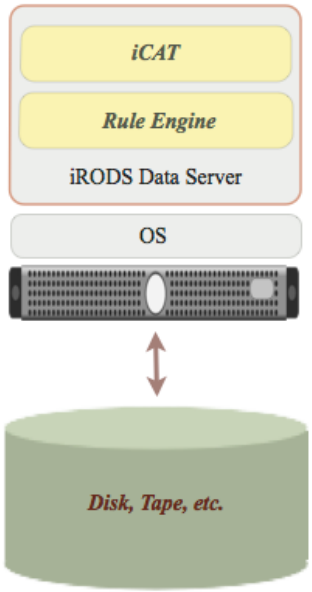
iRODS: integrated Rule-Oriented Data System



Architecture

•Front end layer – provides access interfaces and communication protocols.

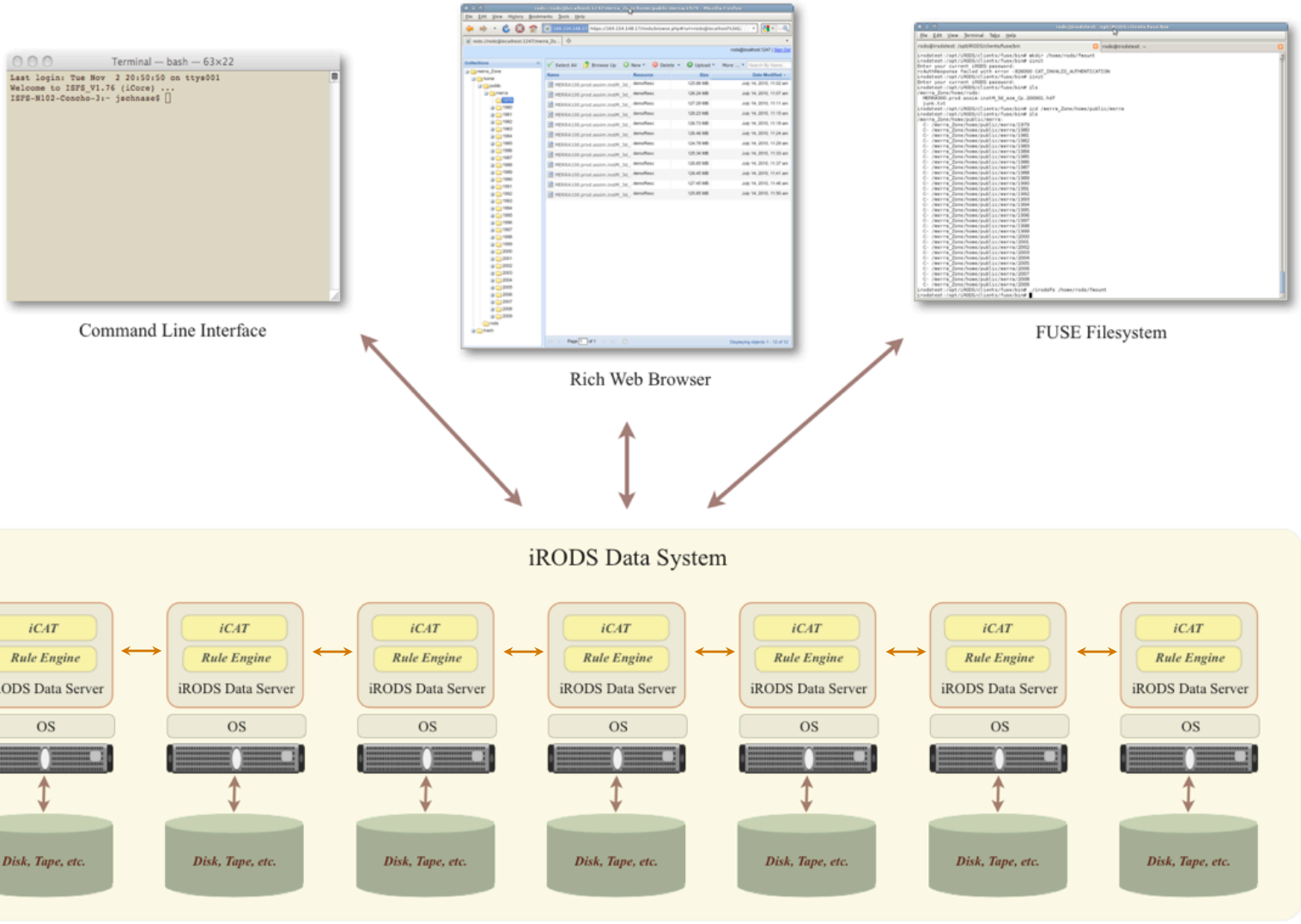
•Back end layer – provides translation drivers for connecting to and communicating with storage systems, such as file systems, tape archives, databases, object-based sensor streams.



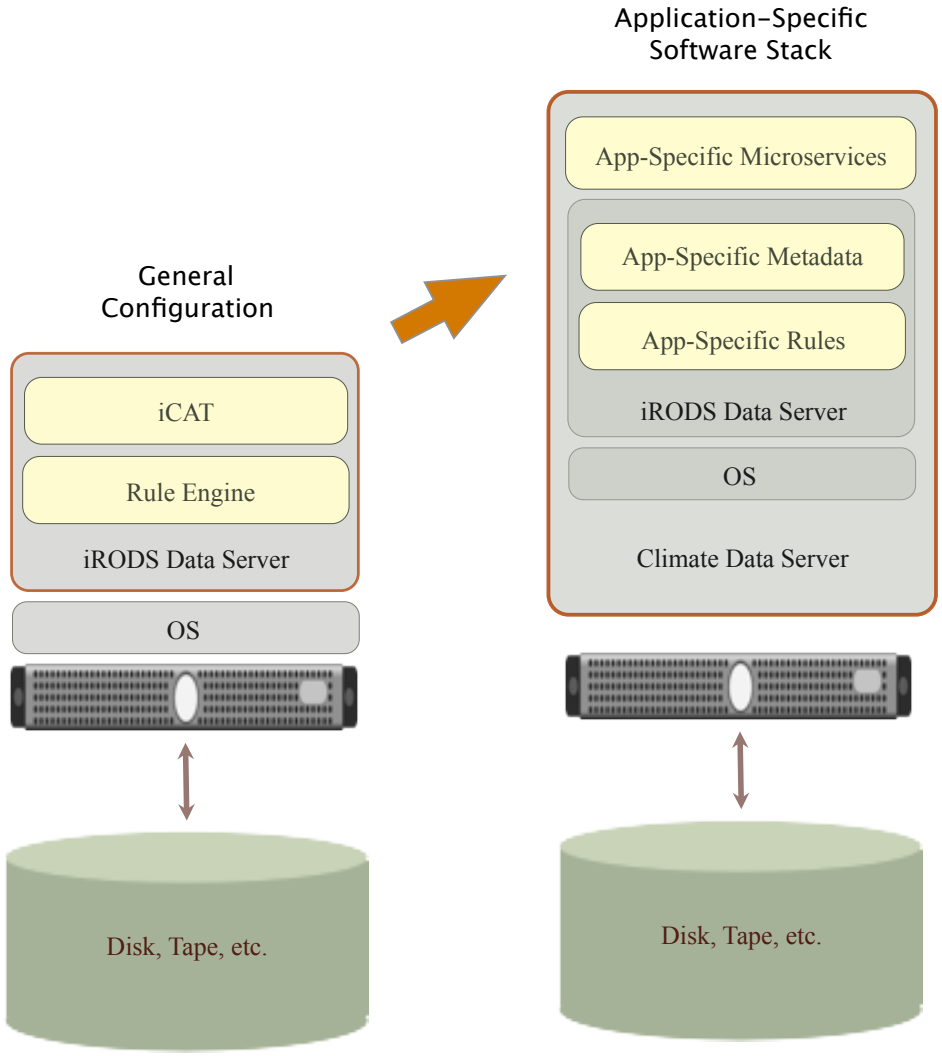
- Middle layer – provides physical transparency by hiding the idiosyncrasies of the client and diver levels.

The middle layer is split between the metadata catalog (iCat) and a rule engine supporting microservices that allow flexibility in defining operational semantics.

iRODS: integrated Rule-Oriented Data System



iRODS-Based Climate Data Server



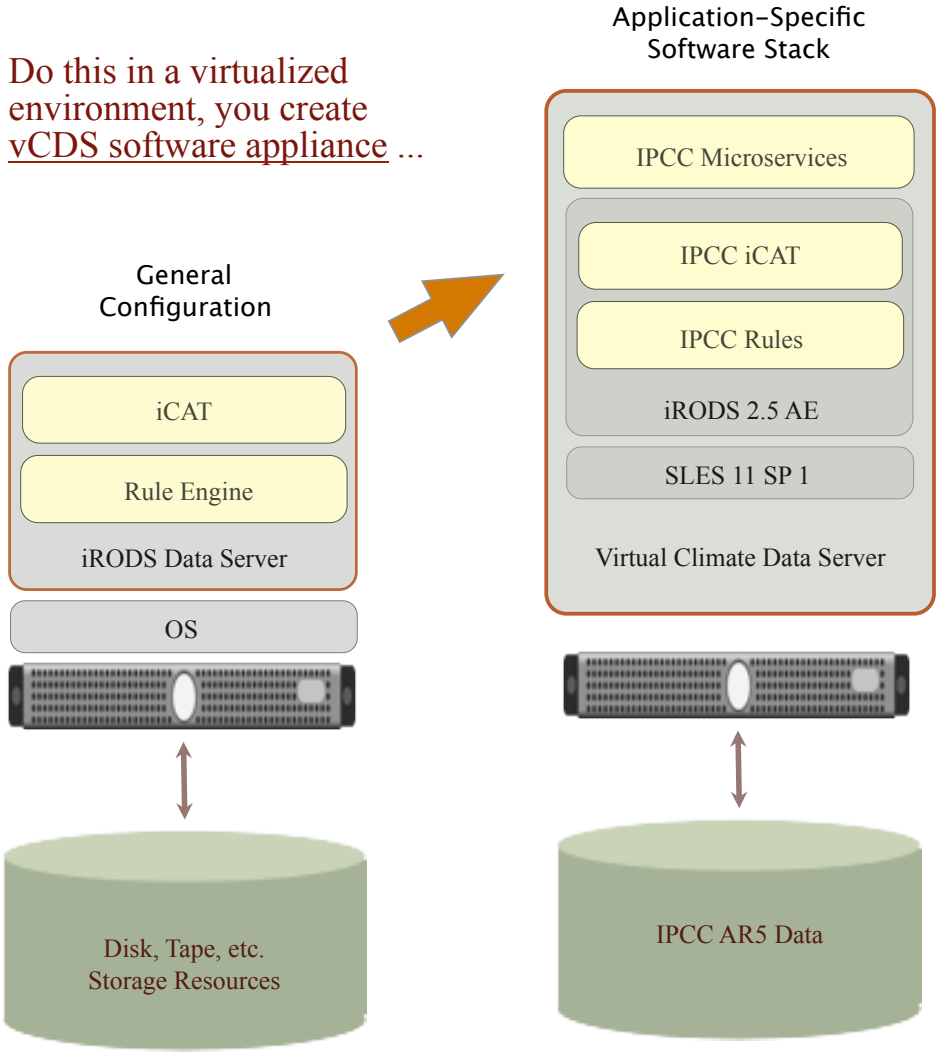
Core Components

- Application-specific microservices
- Application-specific metadata
- Application-specific rules
- A specific release of iRODS
- A specific operating system

Do this in an organized way, you create a CDS software appliance ...

iRODS-Based Virtual Climate Data Server for IPCC Data

Do this in a virtualized environment, you create vCDS software appliance ...



vCDS 1.0 Product Suite

IPCC / NetCDF Module

- iRODS microservices, rules, configuration settings, and utilities required for canonical IPCC/NetCDF CRUD operations ...

Administrative Extensions

- iRODS Postgres extensions and utilities to log system-level object provenance and provide QA for OAIS metadata compliance ...

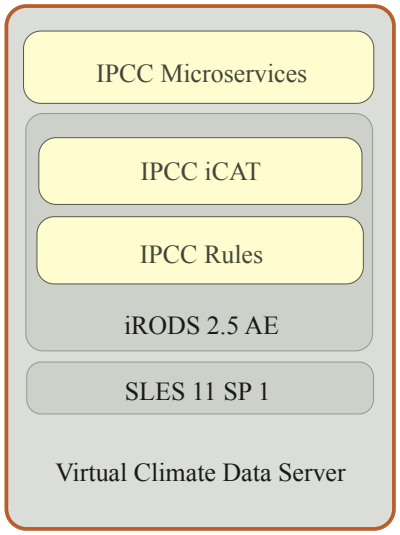
Repetitive Provisioning

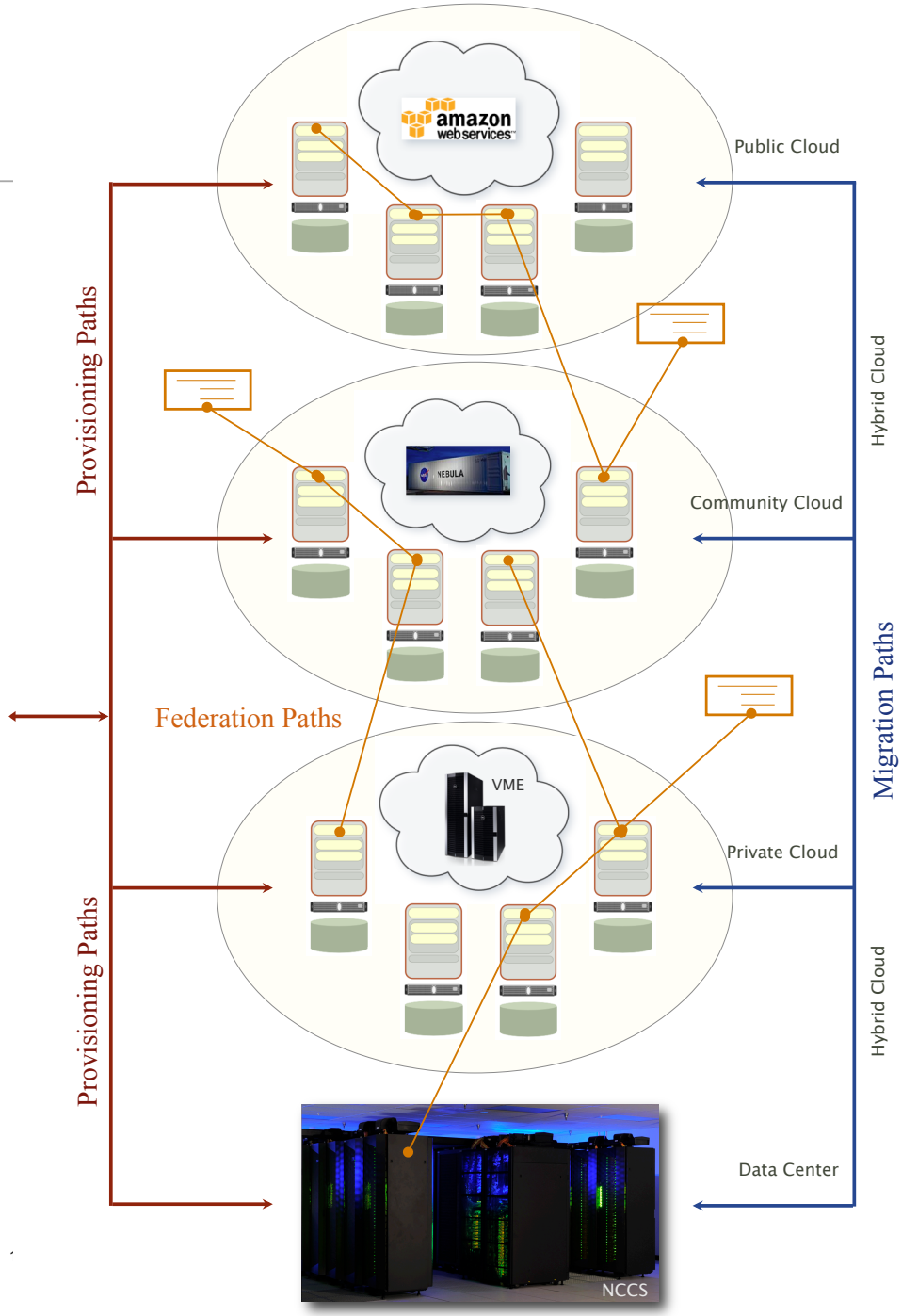
- RPM scripts to build software stacks for the SLES 11 SP1 (IaaS), iRODS AE (PaaS), and CDS/IPCC (SaaS) virtual images ...

Deployment and Distribution

- Product library, documentation, and SLA infrastructure for distribution, deployment, and help desk support ...

vCDS / VaaS Architecture





vCDS / VaaS Architecture

Virtualization is a driving concept

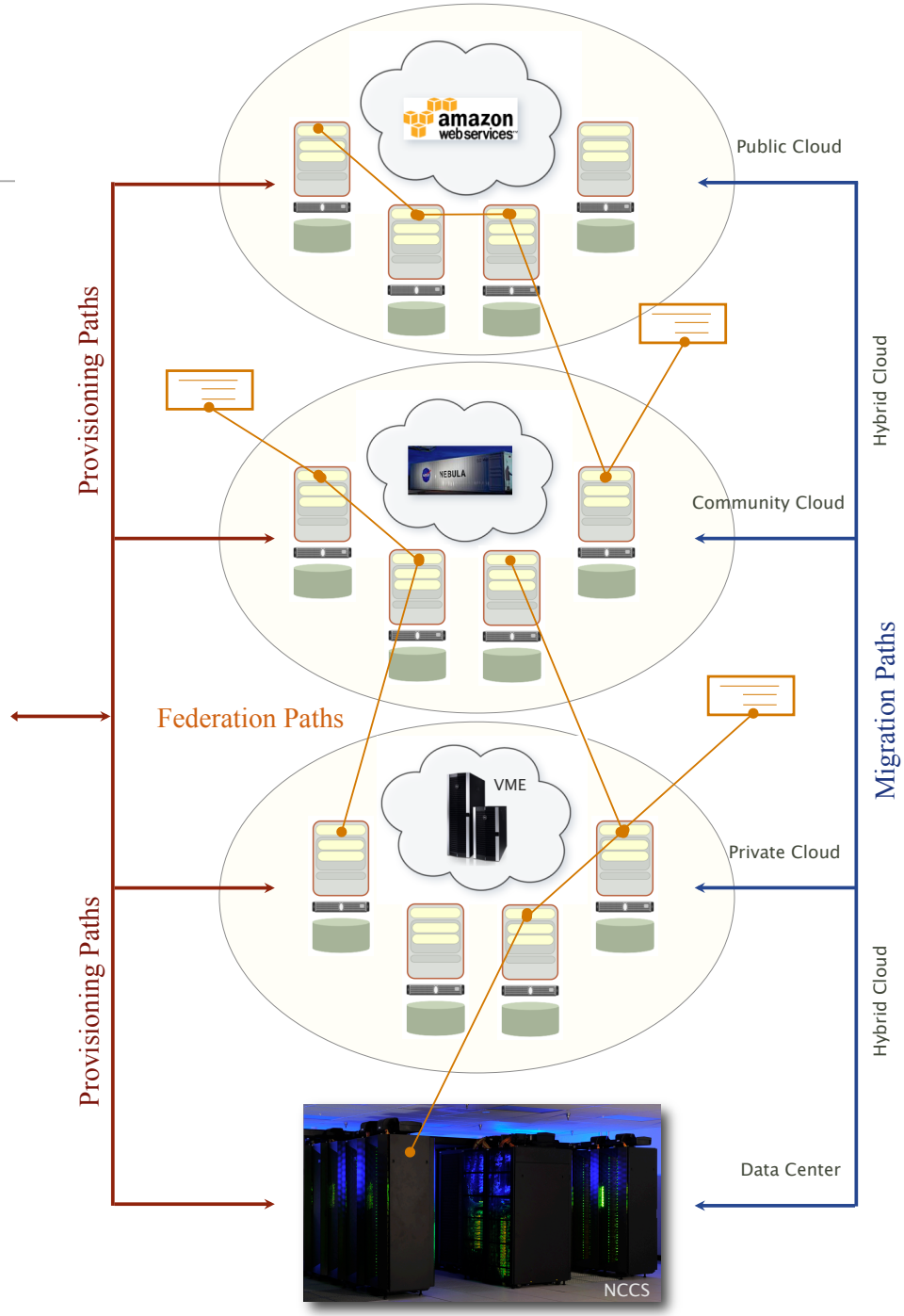
•Provides ready access to a tiered array of services that are flexible, adaptable, scalable, and stageable to NCCS "bricks and mortar" facilities as needed ...

Virtualization-as-a-Service is a huge unmet need in cloud computing

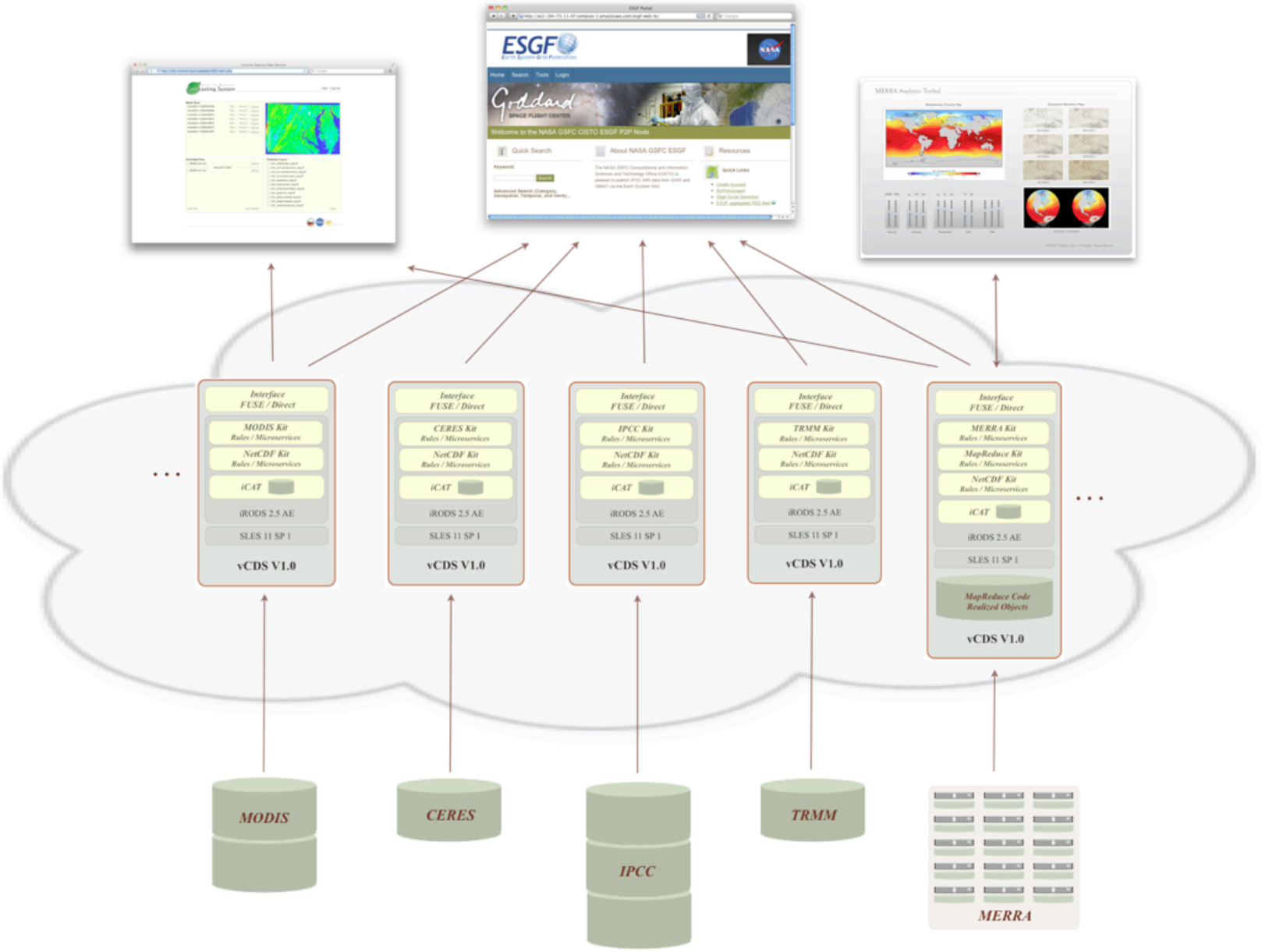
•This approach provides an agile entry point into the NCCS for new customers with data-centric requirements ...

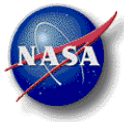
With VaaS, you can distribute or deploy

•If performance is an issue, you can do an old fashioned install. You can distribute images if you're in a virtual world. You can also host images thereby enabling PaaS or SaaS ...



Ecosystem of Managed Collections





iRODS-Based Climate Data Services

John L. Schnase
Office of Computational and Information Science and Technology (Code 606)

NASA Goddard Space Flight Center
September 12, 2012